Learning to be Bayesian without Supervision

EE5111 Estimation Theory final presentation

Rajat V D
EE16B033

Vignesh S
EE16B127

Dhanush Krishna
EE16B009

Table of contents

Recap

Prior-free Bayes estimation

- We rewrite the Bayes estimator (using matrices) without any reference to the prior:

$$\mathbb{E}[X|Y=y] = \frac{(\mathbf{P}_{\mathbf{Y}|\mathbf{X}}\mathbf{X}\mathbf{P}_{\mathbf{Y}|\mathbf{X}}^{-1}\mathbf{P}_{\mathbf{Y}})_y}{(\mathbf{P}_{\mathbf{Y}})_y} = \frac{(\mathbf{L}\mathbf{P}_{\mathbf{Y}})_y}{(\mathbf{P}_{\mathbf{Y}})_y}$$

- With a parametric model f_θ, we minimize the mean square error and simplify it as:

$$\hat{\theta} = \arg\min_\theta \mathbb{E}[(f_\theta(Y) - X)^2] = \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N}\{(\mathbf{f}_\theta)_{Y_i}^2 - 2(\mathbf{L}^T\mathbf{f}_\theta)_{Y_i}\}$$

- For the scalar, zero-mean additive Gaussian noise case ($Y = X + W$), we get the Stein's unbiased risk estimator (SURE):

$$\hat{\theta} = \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N}\{g_\theta(Y_i)^2 + 2\sigma^2 g'_\theta(Y_i)\}$$

- **We found an error in the paper [1] in the final simplified equation for the SURE estimator - the authors missed a factor of 2 in the second term.**

Non-parametric simulations

Beta prior and Binomial likelihood

- In this case, we use a Beta distribution as the prior and Binomial distribution as the likelihood distribution.

$$X \sim \text{Beta}(\alpha, \beta) \quad P_{Y|X} \sim \text{Binomial}(N_x, X)$$

- The sample space of X is uniformly discretized into N_x points. Y correspondingly takes values from 0 to $N_x - 1$.
- After doing this we can directly use the matrix form given below to obtain the estimated X.

$$\mathbb{E}[X|Y = y] = \frac{(\mathbf{P_{Y|X}} \mathbf{X} \mathbf{P_{Y|X}^{-1}} \mathbf{P_Y})_y}{(\mathbf{P_Y})_y}$$

- We compare this alongside MLE, optimal Bayesian and Supervised BLS.

MLE, Bayes optimal and Supervised Estimators

- The MLE estimator maximizes the binomial likelihood:

$$\hat{X}(y) = \arg\max_{\hat{x}} \text{Binomial}(N_x, \hat{x})(y)$$

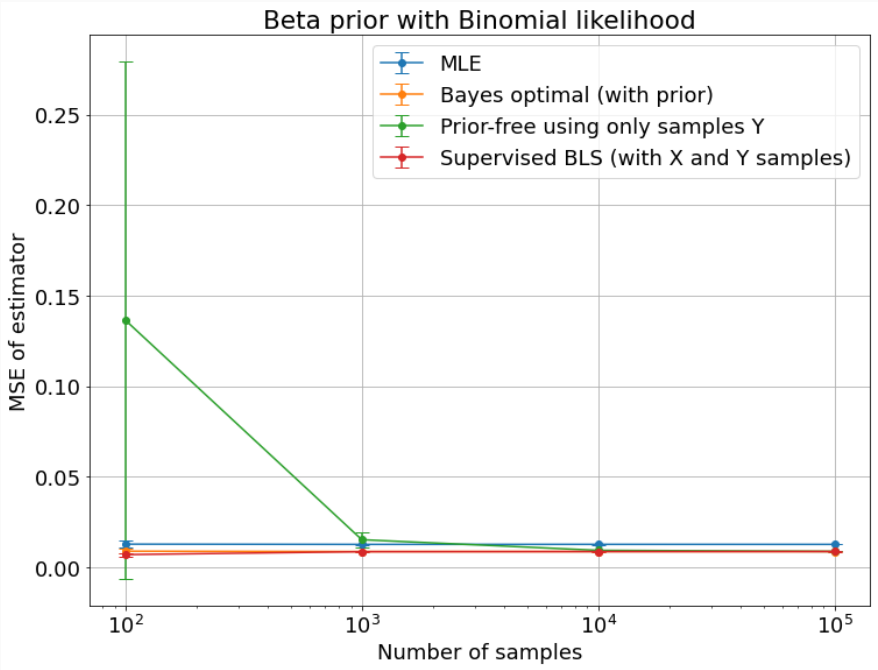
- The Bayes estimator is the MMSE estimator of X given Y. But for a binomial likelihood the beta distribution is the conjugate prior and hence posterior is also beta distribution which is a function of y :

$$\begin{aligned}\hat{X}(y) &= \mathbb{E}[X|Y=y] \\ &= \sum_x x \cdot \text{Beta}(\alpha+y, \beta+n-y)(x) \\ &= \frac{\alpha+y}{\alpha+\beta+n}\end{aligned}$$

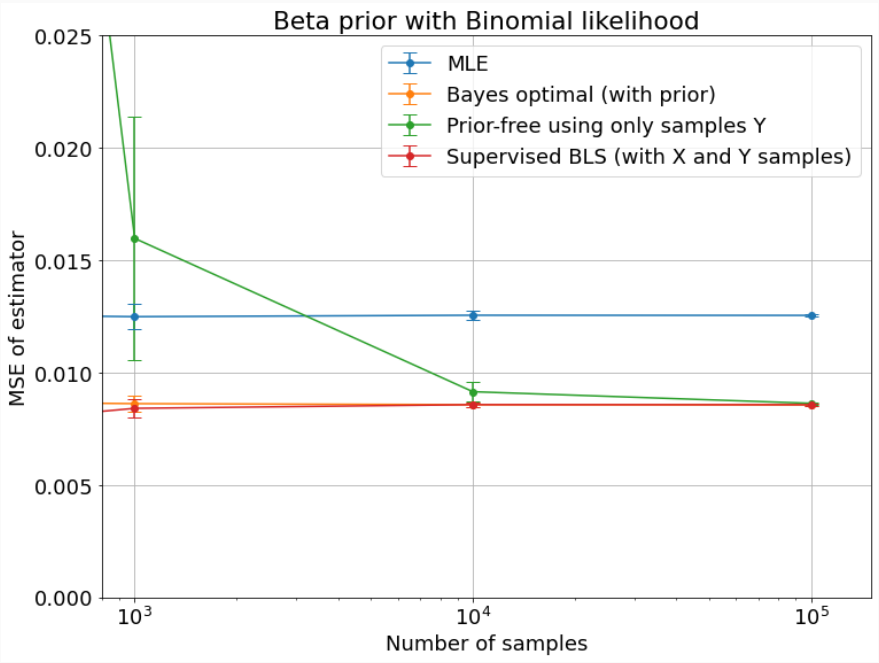
- Finally, we estimate using the supervised non-parametric empirical method as:

$$f(y) = \frac{1}{|\{k : Y_k = y\}|} \sum_{i \in \{k:Y_k=y\}} X_i$$

Simulation Results



(a) Non-parametric estimators.



(b) Zoomed in version of the plot.

Figure 1: Each simulation was run 100 times and the mean plotted with error bars denoting one standard deviation. Observe asymptotic convergence of the prior-free estimator to the Bayes optimal estimator. The supervised estimator converges faster, showing that lack of knowledge of prior leads to **lower sample efficiency**.

Parametric simulations

SURE estimator for additive Gaussian noise

- The SURE estimator is obtained by solving:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \{g_\theta(Y_i)^2 + 2\sigma^2 g'_\theta(Y_i)\}$$

- This is the **empirical prior free estimator** for the additive Gaussian noise model $Y = X + W$ where $W \sim \mathcal{N}(0, \sigma^2)$
- Note that this expression does not depend on the prior of X.
- We compare this estimator with the maximum likelihood and Bayes optimal estimators.

MLE and Bayes optimal estimators

- Since we are dealing with scalar X and Y, the maximum likelihood estimator simplifies:

$$\begin{aligned}\hat{X}_{\text{MLE}}(y) &= \arg\max_x \log P_{Y|X}(y|x) \\ &= \arg\max_x \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right) \\ &= y\end{aligned}$$

- The Bayes optimal estimator is obtained using knowledge of the prior distribution P_X.
- For simplicity, we choose a Gaussian prior, $X \sim \mathcal{N}(\mu_0, \sigma_0)$.
- Since this is a conjugate prior, finding the posterior mean (which is the Bayes optimal estimator) is tractable, and after some algebra we obtain:

$$\mathbb{E}[X|Y=y] = \frac{\sigma_0^2}{\sigma_0^2+\sigma^2} y + \frac{\sigma^2}{\sigma_0^2+\sigma^2}\mu_0$$

- To observe the effect of not having access to the prior, we also compare with a simple supervised linear estimator which is learned from samples (X_i, Y_i).

Simulation results

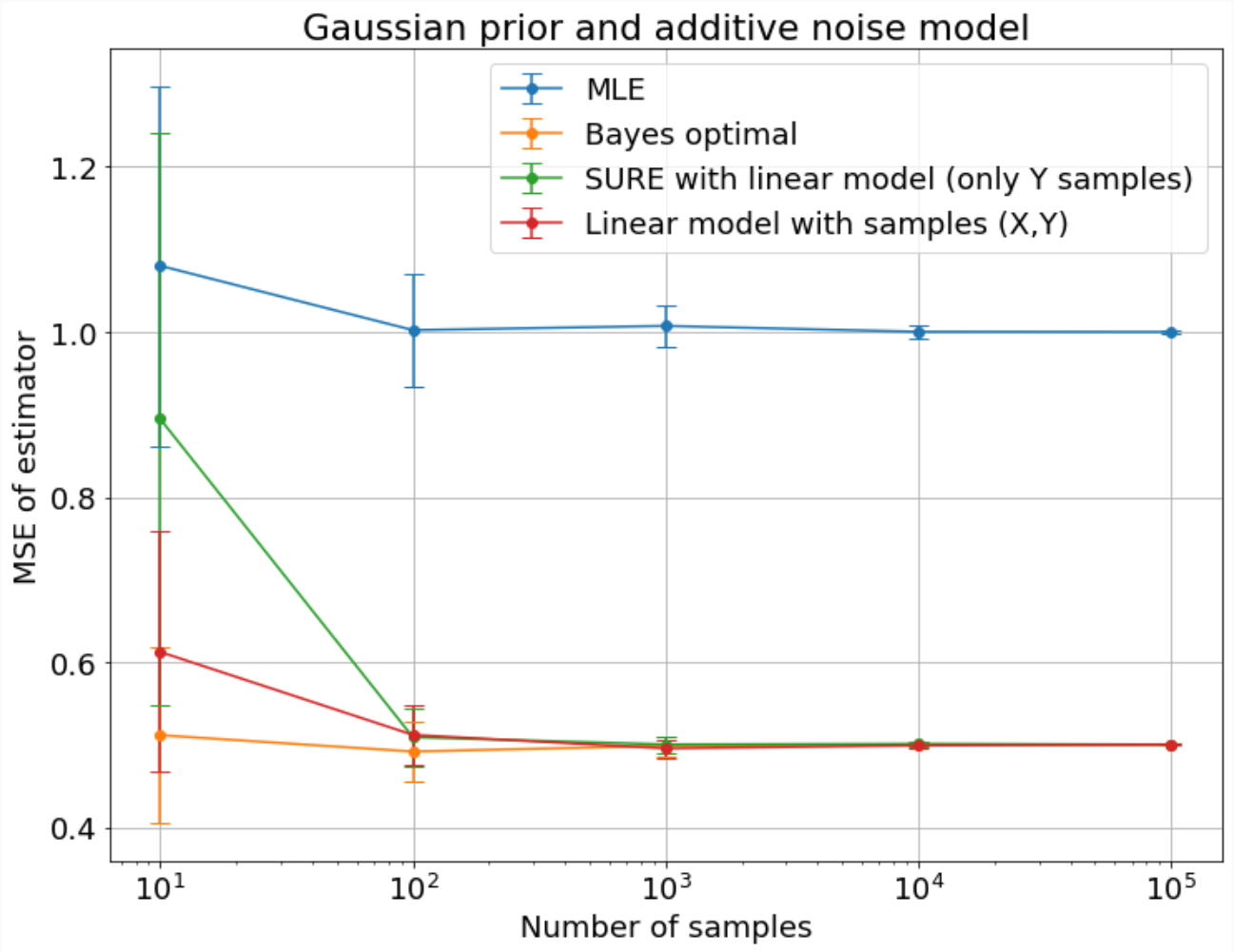


Figure 2: Each simulation was run 100 times and the mean plotted with error bars denoting one standard deviation. Observe asymptotic convergence of the prior-free estimator to the Bayes optimal estimator.

Empirical Bayes and Conjugate Priors

Learning the conjugate prior

- The SURE estimator aims to find the Bayes optimal estimator using only knowledge of the likelihood $P_{Y|X}$ and samples Y_i.
- This setting is also tackled by other methods, including empirical Bayesian methods.
- Empirical Bayesian methods model the prior distribution in terms of **fixed parameters** γ (as opposed to random parameters in the case of heirarchical Bayes).
- To simplify inference, we assume that the prior is a **conjugate prior** (Gaussian in our case).
- Empirical Bayes estimates the parameters γ by **MLE on the marginal likelihood** $P_{Y|\gamma}$.
- Inference is then done using the learned parameters of the prior.

Empirical Bayes for additive Gaussian noise

- For the additive Gaussian noise model, the conjugate prior is also Gaussian $X \sim \mathcal{N}(\mu_\pi, \sigma_\pi^2)$.
- With this Gaussian prior, the observed variable $Y = X + W$ is also Gaussian.

$$E[Y] = \mu_\pi$$

$$\text{Var}(Y) = \sigma_\pi^2 + \sigma^2$$

- We can estimate the mean and variance of the marginal likelilood (P_Y) by MLE:

$$\hat{\mu} = \frac{1}{N} \sum_i Y_i$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_i (Y_i - \hat{\mu})^2$$

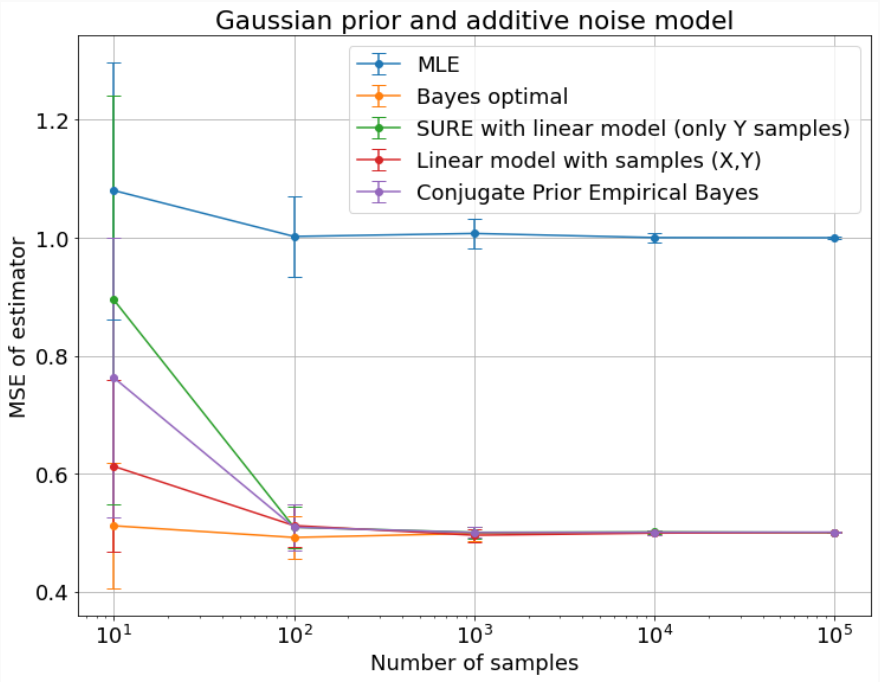
- We can then obtain the prior parameter estimates:

$$\hat{\mu}_\pi = \hat{\mu}$$

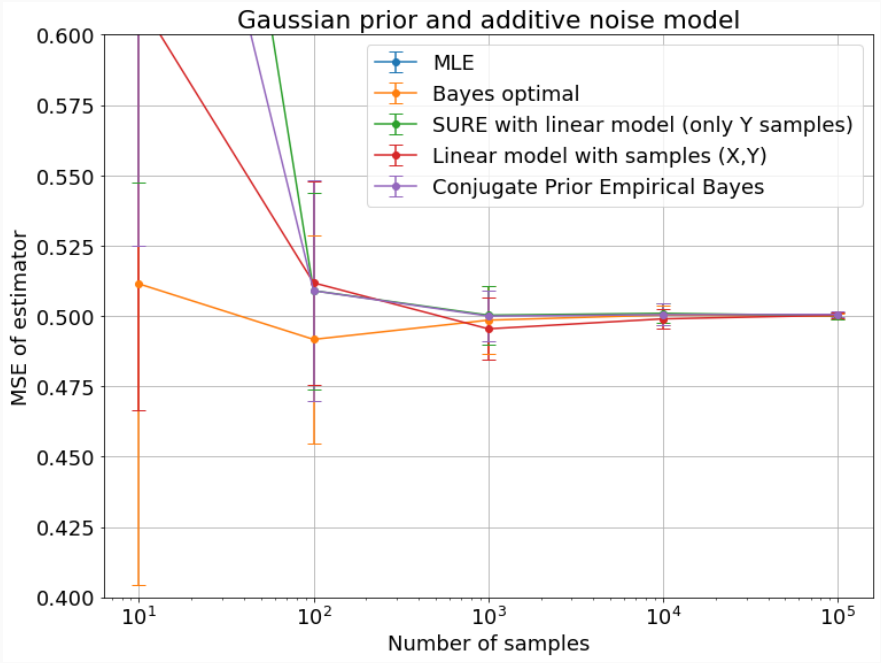
$$\hat{\sigma}_\pi^2 = \hat{\sigma}^2 - \sigma^2$$

- Bayesian inference using this prior is simple because it is a conjugate prior.

Simulation results



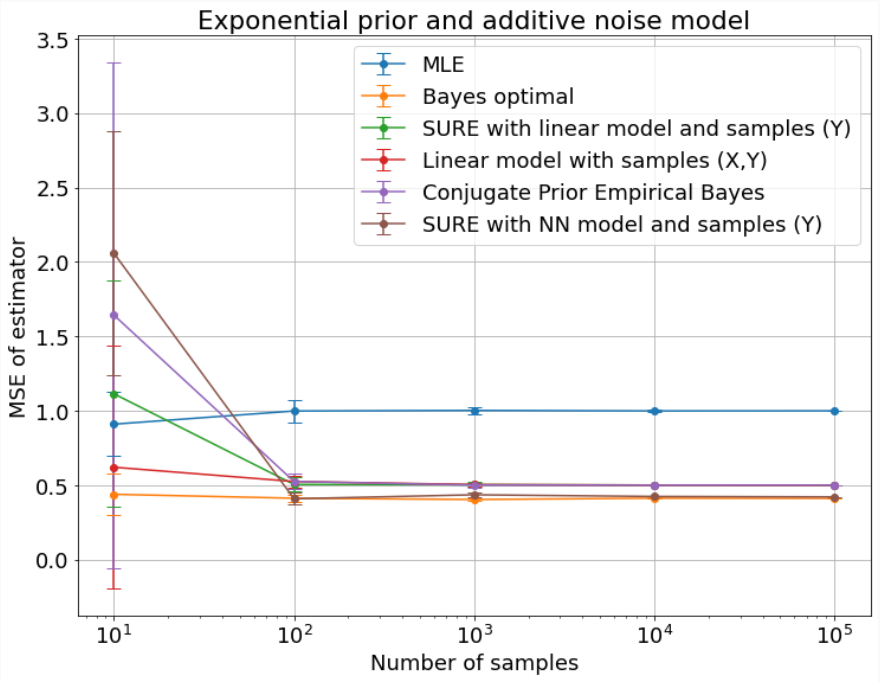
(a) Empirical Bayes using conjugate prior.



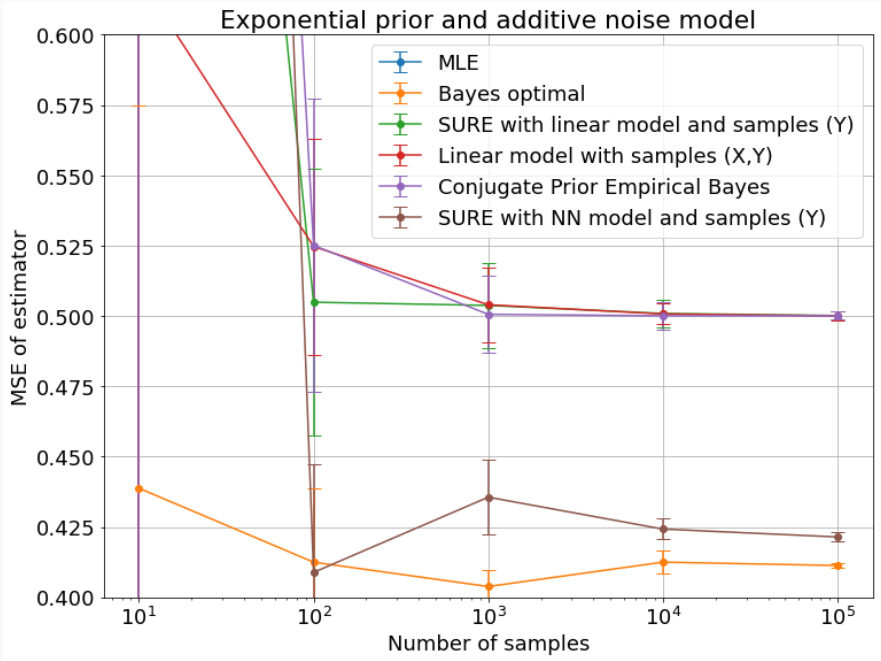
(b) Zoomed in version of the plot.

Figure 3: Observe that Empirical Bayes performs better than the prior-free estimator for small sample sizes. Empirical Bayes also asymptotically converges to the Bayes optimal because the conjugate prior assumption is correct (the actual prior is Gaussian). Note that this might not hold for other priors.

Exponential prior with Neural Network



(a) SURE with a neural network model.



(b) Zoomed in version of the plot.

- Empirical Bayes performs worse for small sample sizes because the conjugate prior assumption is not valid.
- All linear models as well as empirical Bayes converge to the same but sub-optimal estimator.
- By using a **neural network** to parameterize g_θ in the SURE estimator, we have enough capacity to approach much closer the Bayes optimal estimator.

Conclusion

- All prior-based methods perform better than MLE asymptotically, but their performance degrades for low sample counts.
- If we know that the actual prior is conjugate to the likelihood, empirical Bayes is the best approach.
- Although the prior-free method converges asymptotically to the Bayes optimal estimator, it is **less sample efficient** compared to other estimators which have access to the prior (either through samples or direct knowledge of the distribution).
- Additionally, the prior-free SURE estimator only converges if the model has sufficient capacity (like a neural network) for more complex choices of the prior.

References

[1] Martin Raphan and Eero P Simoncelli. Learning to be bayesian without supervision. In *Advances in neural information processing systems*, pages 1145–1152, 2007.